J Chron Dis Vol. 38, No. 10, pp. 811–830, 1985
Printed in Great Britain.

0021-9681/85 $3.00 + 0.00

Variance and Dissent

STATISTICAL AGE–PERIOD–COHORT ANALYSIS: A REVIEW AND CRITIQUE

LAWRENCE L. KUPPER, JOSEPH M. JANIS, AZZA KARMOUS and BERNARD G. GREENBERG

Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, NC 27514, U.S.A.

(*Received* 24 *August* 1984)

Abstract—Descriptive and statistical age–period–cohort (APC) analysis methods have received considerable attention in the literature. The statistical modeling of APC data often involves the popular multiple classification model, a model containing the effects of age groups (rows), periods of observation (columns), and birth cohorts (diagonals of the age-by-period table). The identifiability problem inherent to this model is discussed, and its adverse effects on the results of APC modeling exercises are illustrated numerically. Potential problems attendant with the use of two-factor models are described, and other possible modeling approaches currently in use are discussed. Interpretational limitations due to certain innate characteristics of typical APC data sets are also detailed. Given all the documented potential sources for error, the current state-of-the-art regarding the statistical modeling of APC data should be considered to be at an early stage of development.

1. INTRODUCTION

AGE–PERIOD–COHORT (APC) analysis has been a popular epidemiologic tool since Frost [1] employed it in his classical study of tuberculosis. The procedure he developed is primarily descriptive, with graphs used to examine patterns in disease rates over time.

In the past few years, however, many studies appearing in the epidemiologic literature have utilized a more analytical approach to the treatment of APC data, such an approach involving the statistical fitting of regression models designed to quantify the *separate* effects of the three factors age, period, and cohort. These investigations have encompassed several different diseases such as breast cancer [2, 3], colon cancer [4], cancer of the cervix [5], prostate cancer [6, 7] bladder cancer [8, 9], and lung cancer [9–13]. Each study has adopted a regression analysis approach to the treatment of incidence or mortality data; typically, a three-factor model (age at occurrence of disease or death, time of occurrence of disease or death, and birth cohort), a two-factor model (usually, age at occurrence and birth cohort), or some modification of these two models has been employed. Many of these types of studies have resulted in epidemiologic statements being made about the etiology of the diseases under investigation.

To an epidemiologist who is not thoroughly acquainted with the specifics of various statistical modeling approaches currently used to analyze APC data, the results and ultimate utility of such quantitative analyses can be quite hard to evaluate. In particular, some reasonable questions that a health researcher might ask about such modeling exercises are the following: How much credence can be given to the results of such statistical analyses? What are the possible sources of error associated with the use of various regression procedures for modeling APC data? Do such sophisticated modeling procedures truly provide interpretational advantages over traditional graphical approaches?

Our purpose in writing this paper is to review and critique the general area of age–period–cohort analysis and to discuss and illustrate some of the important limitations of popular statistical modeling approaches for analyzing APC data. We will argue that any interpretations regarding patterns in age, period, and cohort effects based on the use of such modeling procedures must be made with a great deal of caution. Furthermore, we will stress that any statistical modeling of APC data should be carried out in conjunction with a detailed descriptive analysis such as discussed by Kleinbaum *et al.* [14] and Glenn [15]. In this paper, we will deal with statistical APC methodology only as it applies to epidemiologic data, although APC analysis has been an integral part of many disciplines such as sociology, demography, developmental psychology, political science, and economics [16–26].

2. DESCRIPTIVE AGE–PERIOD–COHORT ANALYSIS

A descriptive APC analysis of epidemiologic data first assembles disease mortality or morbidity rates in a two-way table with, say, the rows representing categories of age at occurrence and the columns defining categories of year of occurrence. In general notation, for the ith of a age groups and the jth of p periods, we will let $\hat{R}_{ij} = O_{ij}/N_{ij}$ denote the observed rate in the (i, j)th cell of such a table, where O_{ij} is the observed number of deaths or illnesses and N_{ij} is the number of person-years at risk.

An example of such a data layout is presented in Table 1, which gives lung cancer mortality rates for United States white males. These rates are based on 5-year age intervals and 5-year period intervals. (Although age and period intervals of the same width are not required, we will make this generally unrestrictive assumption to avoid certain esoteric mathematical complexities [7, 17].)

The diagonals of this Table (going from upper left to lower right) define the lung cancer mortality rate patterns for successive groups of U.S. white males who were born together and hence age together. In Table 1, these so-called *birth cohorts* extend over 9-year intervals, and each such birth cohort is typically identified by its *central* birth year. For

TABLE 1. LUNG CANCER MORTALITY RATES* PER 100,000 ($10^5\hat{R}_{ij}$) AND AVERAGE NUMBERS OF LUNG CANCER DEATHS PER YEAR (O_{ij})† FOR U.S. WHITE MALES BY YEAR (1931–1975) AND BY AGE (30–84), BOTH IN FIVE-YEAR INTERVALS

Age group		Calender period								
		1931–35	1936–40	1941–43	1946–50	1951–55	1956–60	1961–65	1966–70	1971–75
30–34	Rate	1.08	1.43	1.64	1.58	1.52	1.96	2.16	2.14	1.74
	Deaths	45	64	78	79	80	104	105	103	97
35–39	Rate	2.38	3.06	3.66	4.10	4.32	5.26	6.35	7.25	7.21
	Deaths	98	127	163	197	219	282	334	357	350
40–44	Rate	4.51	6.78	8.46	10.05	11.75	13.40	16.07	19.60	19.87
	Deaths	180	271	352	447	559	673	842	1041	990
45–49	Rate	8.19	13.72	17.62	22.31	27.46	30.75	36.16	41.06	47.51
	Deaths	288	521	695	907	1197	1453	1755	2132	2458
50–54	Rate	13.04	21.76	30.35	43.32	53.82	64.29	72.11	81.43	86.42
	Deaths	391	720	1082	1609	2054	2656	3197	3861	4407
55–59	Rate	16.14	30.07	47.06	68.18	88.37	108.86	121.86	139.46	151.86
	Deaths	397	812	1391	2201	3050	3939	4788	5961	6743
60–64	Rate	19.61	34.75	53.65	87.38	120.90	158.32	189.36	218.81	238.48
	Deaths	381	755	1295	2365	3614	4933	5943	7640	9129
65–69	Rate	20.73	38.22	54.29	87.47	136.24	189.11	240.25	289.07	325.13
	Deaths	301	635	1026	1867	3265	4937	6361	7973	9922
70–74	Rate	20.77	32.98	50.36	85.62	128.85	184.99	245.18	322.38	390.94
	Deaths	206	370	648	1247	2153	3563	5132	6824	8447
75–79	Rate	19.09	33.59	44.96	77.73	115.21	160.81	224.70	318.68	415.37
	Deaths	110	218	334	680	1186	1906	2983	4591	5976
80–84	Rate	12.47	23.49	31.46	64.03	96.88	132.16	179.54	260.91	360.16
	Deaths	36	76	114	281	495	788	1223	2066	3083

*The National Center for Health Statistics (NCHS) supplied the lung cancer mortality rates for the years 1950 through 1975. Sources for the remaining data are: U.S. Department of Commerce, Bureau of the Census: Estimates of the population of the United States by age, sex, and race. **Current Population Reports**, Series P-25, No. 311, 1931–1949; and U.S. Department of Health, Education and Welfare, National Center for Health Statistics: **Vital Statistics of the U.S.**, Annual Mortality Volumes, 1931–1949. Further details about the data can be found in: Janis JM: A Descriptive and Statistical Methodology for Age–Period–Cohort Analysis with Application to Lung Cancer. *Unpublished doctoral dissertation. Department of Biostatistics, University of North Carolina, 1981.*

†$N_{ij} = O_{ij}/\hat{R}_{ij}$.

example, the "1901 birth cohort" contains those males born during the 9-year period 1897 through 1905; in Table 1, the diagonal containing the nine rates 1.08×10^{-5}, 3.06×10^{-5}, $8.46 \times 10^{-5}, \ldots, 289.07 \times 10^{-5}$, 390.94×10^{-5} gives the mortality rate pattern for this particular birth cohort. The diagonal just below this one contains the nine rates pertaining to the "1896 birth cohort" (i.e. those males born between 1892 and 1900). The reader should be aware of the fact that birth cohorts will "overlap" (i.e. will have certain years of birth in common) when defined in this manner; such overlap is typically ignored when fitting statistical models to APC data.

Another important characteristic of this correspondence between birth cohorts and the diagonals of an age-by-period two-way data layout is that birth cohorts corresponding to diagonals at the extremes of the Table will involve very few data points (e.g. the 1851 and 1941 birth cohorts contain only one observation each). We will say more later about the effects of varying birth cohort size upon the interpretation of estimated birth cohort effects.

In a purely descriptive APC analysis, the rates (or transformations thereof) are plotted in various ways as a function of the age, period, and cohort groupings, one such way being that illustrated in Fig. 1. Here, the data in Table 1 are plotted with the rate per 100,000 (i.e. $10^5 \hat{R}_{ij}$) as the ordinate and the age at death category as the abscissa (each age at death category being indexed by the first year of the 5-year age interval). Successive periods of death are shown by solid lines, and successive birth cohorts by dashed lines. (To avoid confusing overlaps, some of the birth cohorts are not shown.)

Although graphs such as Fig. 1 are helpful in obtaining general qualitative impressions about age, period, and cohort rate patterns, they have certain major limitations. For example, in Fig. 1, the pattern in the lung cancer mortality rate curve for the 1961–65 period can be seen to increase steadily from age 30 up to about age 68, before starting to fall off. Note, also, that this curve cuts across a number of birth cohort curves (e.g. 1881, 1886, 1891, etc.). Thus, the shape of this period curve is affected both by varying age effects and by varying cohort effects. Furthermore, a *quantitative* assessment of the way in which these age and cohort effects operate to influence the shape of this period curve cannot be obtained by a simple visual examination of graphs like Fig. 1. Such quantification can only be achieved via the use of *valid* statistical modeling procedures.

3. STATISTICAL AGE-PERIOD-COHORT ANALYSIS: GENERAL CONSIDERATIONS

Greenberg *et al.* [27] first recognized the limitations of a descriptive analysis, and they proposed the use of a three-factor, analysis of variance-type model to quantify the "separate" effects of the (categorized) age, period, and cohort variables. In this paper, we will focus primarily on the so-called APC multiple classification model, a model which has been the most often used in practice, the most thoroughly studied in theory, and the most controversial in character.

The APC multiple classification model has the specific structure

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{a-i+j} + \varepsilon_{ij}, \tag{1}$$

$i = 1, 2, \ldots, a$ and $j = 1, 2, \ldots, p$; here, $Y_{ij} = f(\hat{R}_{ij}) = f(O_{ij}/N_{ij})$ represents some function of the observed rate $\hat{R}_{ij}$, μ is the overall mean, α_i is the *fixed* effect of the ith age category, β_j is the *fixed* effect of the jth period category, and γ_{a-i+j} is the *fixed* cohort effect associated with the ith age category and the jth period category. Note that there are $(a + p - 1)$ distinct cohort effects defined by model (1) as i ranges from 1 to a and j from 1 to p. The only random component on the right-hand side of equation (1) is ε_{ij}, which is assumed to have mean (or expected value) $\mathcal{E}(\varepsilon_{ij}) = 0$. The variance and other distributional properties of ε_{ij} are tied to the assumptions made about the stochastic nature of Y_{ij} and hence of $\hat{R}_{ij}$ (e.g. O_{ij} is often assumed to be Poisson and N_{ij} to be non-random).

For illustrative purposes, the special case $a = 3$, $p = 4$ is diagrammed in Table 2, below, with a typical cell entry being $\mathcal{E}(Y_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{a-i+j}$.

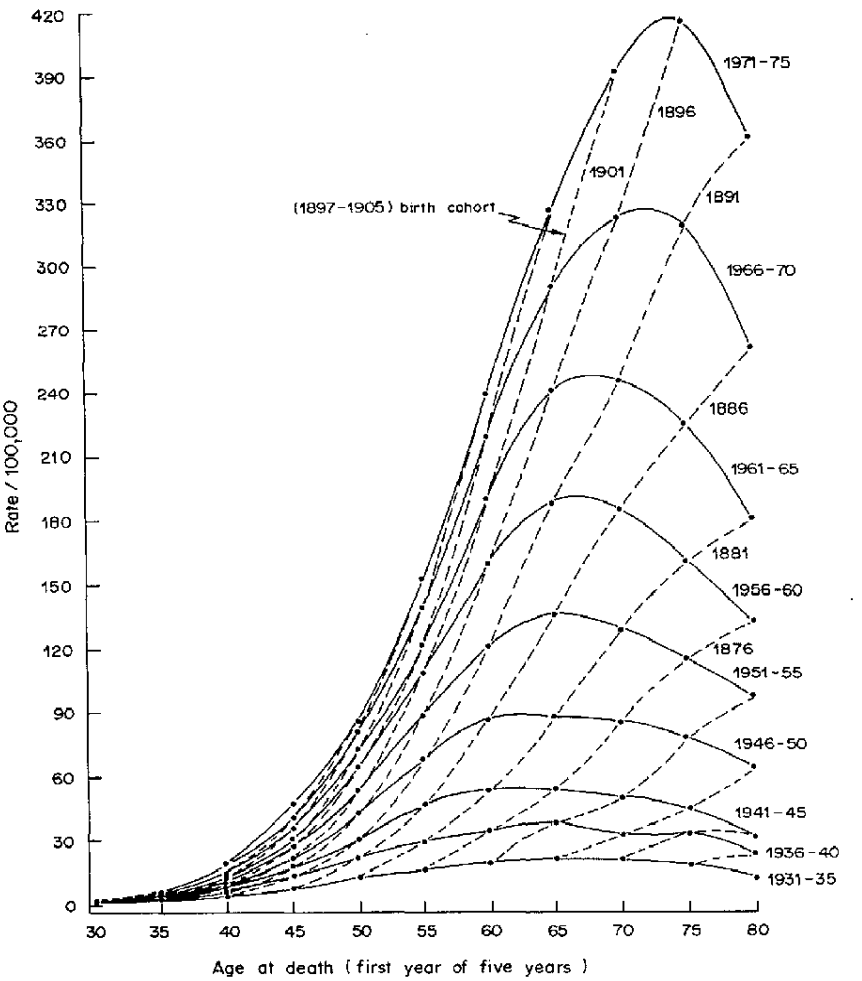


FIG. 1. U. S. white male lung cancer mortality rates per 100,000 (see Table 1) by age at death, period of death, and birth cohort.

From the structure of model (1) and from an inspection of Table 2, it can be seen that the APC multiple classification model (1), as with more standard two-way ANOVA-type models, assumes that the age effects (α_i's) are constant along rows, that the period effects (β_j's) are constant down columns, and that the cohort effects (γ_{a-i+j}'s) are constant along the diagonals. Of course, model (1) is atypical due to the non-standard structure of the "cohort" term, a structure which we will interpret momentarily as a special form of age-by-period interaction effect.

Model (1) is most appropriate for detecting subtle and unexpected patterns in age, period, and cohort effects since it does not assume any *a priori* specific functional relationships to which such patterns must conform. Such *a priori* structural assumptions (e.g. like requiring that the age effects follow some sort of polynomial curve) are generally hard to defend, and can actually restrict the sensitivity of model (1) for detecting important effects in the data.

Without going into the mathematical details, it is important to emphasize that the cohort effects in model (1) characterize a very specific form of interaction between the (categorical) age and period variables. For a detailed mathematical discussion regarding this concept,

TABLE 2. A DIAGRAMMATIC REPRESENTATION OF MODEL (1) FOR THE SPECIAL CASE $a = 3$, $p = 4$

		Period group (j)			
		$j=1$	$j=2$	$j=3$	$j=4$
Age Group (i)	$i=1$	$\mu+\alpha_1+\beta_1+\gamma_3$	$\mu+\alpha_1+\beta_2+\gamma_4$	$\mu+\alpha_1+\beta_3+\gamma_5$	$\mu+\alpha_1+\beta_4+\gamma_6$
	$i=2$	$\mu+\alpha_2+\beta_1+\gamma_2$	$\mu+\alpha_2+\beta_2+\gamma_3$	$\mu+\alpha_2+\beta_3+\gamma_4$	$\mu+\alpha_2+\beta_4+\gamma_5$
	$i=3$	$\mu+\alpha_3+\beta_1+\gamma_1$	$\mu+\alpha_3+\beta_2+\gamma_2$	$\mu+\alpha_3+\beta_3+\gamma_3$	$\mu+\alpha_3+\beta_4+\gamma_4$

the reader is encouraged to consult the paper by Kupper *et al.* [28]. For now, it is sufficient for the reader to appreciate the fact that the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \tag{2}$$

which contains no cohort effects, is a no-interaction model in the sense that the expected response in cell (i, j) is determined as a function of i and j *solely* by the *marginal* (or separate) effects of row i and column j, and not also by an effect (such as γ_{a-i+j}) which is a function of *both* i and j (i.e. is cell-specific).

To pursue this notion further in the context of a commonly used epidemiologic data analysis procedure, let us contrast models (1) and (2) with regard to the use of age adjustment by the direct method when assessing trends in incidence or mortality rates over time. In particular, let

$$(sY)_j = \sum_{i=1}^{a} W_i Y_{ij}$$

denote the directly age-standardized rate function for the jth period; the $\{W_i\}$ are weights

$$\text{(satisfying } \sum_{i=1}^{a} W_i = 1\text{)}$$

which vary only with i and which are usually based on the age distribution in some external standard population (e.g. the 1970 U.S. population).

Under the no-interaction model (2), it follows, for $j \neq j'$, that

$$\mathcal{E}[(sY)_j - (sY)_{j'}] = (\beta_j - \beta_{j'}),$$

so that a plot of the $\{(sY)_j\}$ vs j would, as desired, correctly display patterns in the period effects (ignoring, of course, the vagaries of sampling error). However, under model (1), for $j \neq j'$,

$$\mathcal{E}[(sY)_j - (sY)_{j'}] = (\beta_j - \beta_{j'}) + \sum_{i=1}^{a} W_i(\gamma_{a-i+j} - \gamma_{a-i+j'}),$$

so that plots of the $\{(sY)_j\}$ vs j would unfortunately reflect a complex and difficult to interpret mixture of varying period and cohort effects.

Deviations from model (2) can sometimes be detected by making plots of the $\{Y_{ij}\}$ vs j for each i (called age-specific plots). A lack of "parallelism" among these curves may suggest the influence of important "interaction" effects. Indeed, epidemiologists have often argued that the presence of such non-parallelism provides some evidence that real birth cohort effects are actually operating [29, 30]. As an example, Figs 2 and 3 give the age group-specific curves for the lung cancer data in Table 1; Fig. 2 uses $10^5\hat{R}_{ij}$ itself, while Fig. 3 employs $\ln(10^5\hat{R}_{ij})$. There is evidence of non-parallelism in both figures, although the use of the log transformation in Fig. 3 tends to dampen the effect somewhat.

In this regard, an important point needs to be made about the choice of the function $f(\hat{R}_{ij})$ and its impact upon the interpretation of estimated age, period, and cohort effects. Figures 2 and 3 help somewhat to illustrate graphically that trends in data (and hence estimated effects based on the modeling of such trends) are affected by the choice of dependent variable $Y_{ij} = f(\hat{R}_{ij})$ used for analysis. Numerous choices for $f(\hat{R}_{ij})$ have been considered in the literature, some of the more popular ones being $f(\hat{R}_{ij}) = \hat{R}_{ij}$, $f(\hat{R}_{ij}) = \ln(\hat{R}_{ij})$ when $\hat{R}_{ij} > 0$, $f(\hat{R}_{ij}) = \ln(1 + \hat{R}_{ij})$ to permit consideration of observed rates equal to zero, and the logit transformation $f(\hat{R}_{ij}) = \ln[\hat{R}_{ij}/(1 - \hat{R}_{ij})]$ for rates that have been suitably scaled to lie between 0 and 1 in value.

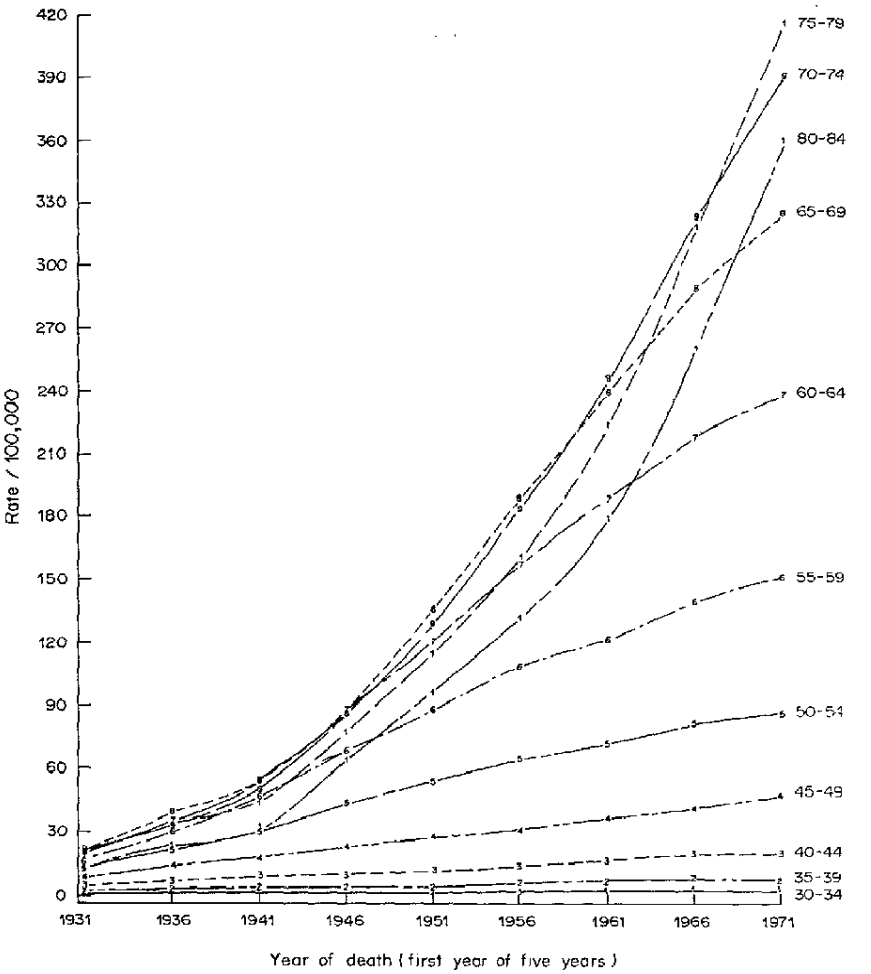
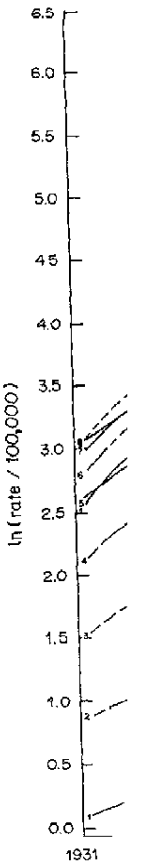


FIG. 2. Age-specific plots of rate per 100,000; data obtained from Table 1.

The reason that the choice for $f(\hat{R}_{ij})$ has such an important role in APC analysis is that interaction quantification (and hence cohort effect estimation) is very much dependent on the scale used to measure such interaction. Kleinbaum *et al.* ([14], Chapter 19) provide a detailed discussion of the controversy in the biostatistical and epidemiological literature about methodological issues of interaction assessment and interpretation in health-related data. In particular, these authors emphasize the inadvisability of using summary rates [e.g. like the $(sY)_j$'s] where there is evidence of interaction. Freeman and Holdford [31] advocate looking for a transformation $f(\hat{R}_{ij})$ which will eliminate or at least suppress strong interaction effects in order to justify the use of "smoothed" summary indices. However, such a search may not be desirable in situations where the presence of certain interaction effects can suggest important causal mechanisms and so merit in-depth study.

The fact that interaction quantification is so very much a model-dependent procedure often creates unresolvable interpretational problems regarding estimated interaction effects, a prime example being patterns in estimated cohort effects based on fitting models like equation (1) to APC data. Since cohort effects are age-by-period interaction effects of a very specific structure, it is clear that the choice of the function $Y_{ij} = f(\hat{R}_{ij})$ will significantly affect the patterns in estimated cohort effects based on modeling Y_{ij} as a function of the variables age, period, and cohort. In fact, it is not unreasonable to suspect



FIG

that, for a give
to the (possibly
in that data se
analysis depen

Having defi
the APC mu
accurately as
of estimates h
plots of the c
etiologically
cohort paran
that the esti
covariance) f
for the peri
whereby "p
obtained, ha
equation (1)

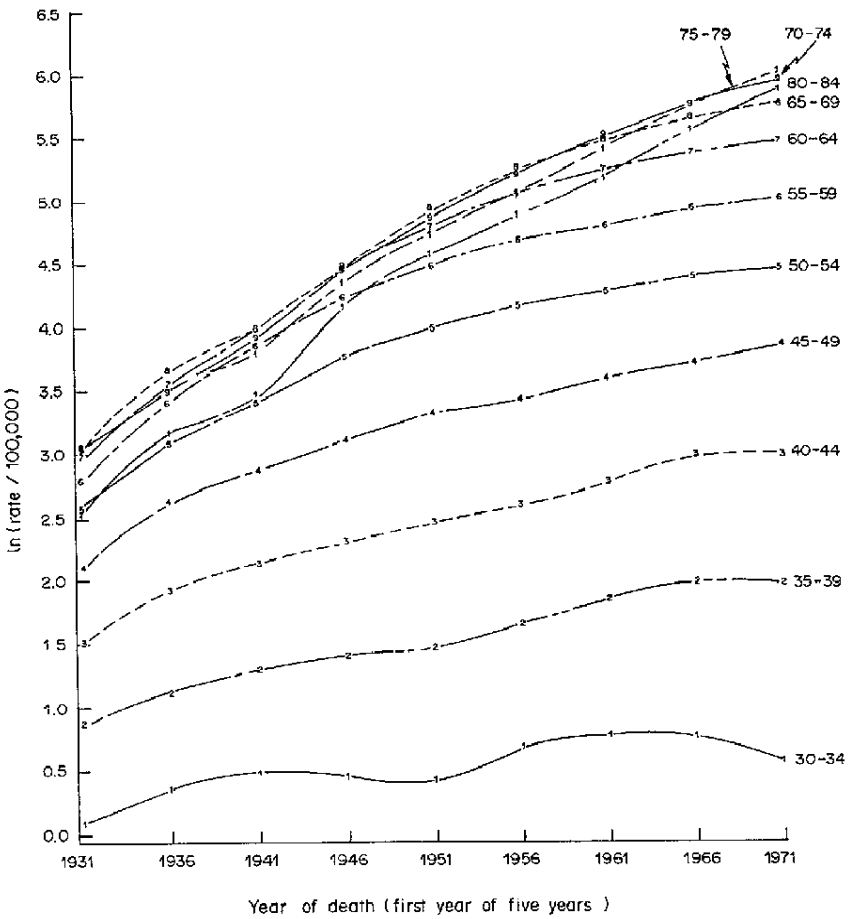


FIG. 3. Age-specific plots of ln (rate per 100,000); data obtained from Table 1.

that, for a given APC data set, a transformation $f(\hat{R}_{ij})$ might be found which would lead to the (possibly inappropriate) conclusion that no important cohort effects are operating in that data set. The reader should bear in mind that the results of any statistical APC analysis depend intimately on the choice of the function $Y_{ij} = f(\hat{R}_{ij})$ which is to be modeled.

4. ESTIMATION OF MODEL (1) PARAMETERS

Having defined (and accepted as a reasonable approximation to the true state of nature) the APC multiple classification model (1), the statistical goal is then to estimate as accurately as possible the age, period, and cohort parameters in that model. Once a set of estimates has been obtained by some procedure, it is standard practice to make separate plots of the estimated age, period, and cohort effects, and then to attempt to "interpret" etiologically any interesting patterns which may emerge. Since these age, period, and cohort parameters appear together in the same model, it is natural to argue, for example, that the estimated age parameters have been "adjusted" (as in the spirit of analysis of covariance) for the effects of the period and cohort factors, with similar claims being made for the period and cohort effect estimates. This statistical adjustment phenomenon, whereby "pure" patterns in estimated age, period, and cohort effects can supposedly be obtained, has served as the primary impetus for the popularity of regression models like equation (1) when analyzing APC data.

As we have emphasized earlier, one major concern associated with modeling APC data using equation (1) is the choice for the transformation $Y_{ij} = f(\hat{R}_{ij})$. However, even if an appropriate specification for $f(\hat{R}_{ij})$ can be made, the utility of APC analysis is additionally seriously threatened by the so-called "identification problem".

4.1. The identification problem: the "bête noire" of APC researchers

The identification problem in APC analysis pertains to the fact that the factors age, period, and cohort are mathematically related. When these factors are treated as continuous variables, this mathematical relationship (when mortality is the endpoint) is simply "(year of birth) + (age at death) = (year of death)." For example, a person who is born in 1900 will have a lifespan of 50 years if he or she dies in 1950. The exact structure of the mathematical dependency among age, period, and cohort variables which have been categorized as in model (1) is harder to quantify; however, such a quantification is provided by Theorem 3.1 in Kupper *et al.* [28].

To appreciate more fully why the presence of such a linear dependency causes a problem when analyzing APC data (like that in Table 1) with the multiple classification model (1), it is helpful to work with an equivalent form of this model. In particular, let us define the parametric means

$$\bar{\alpha} = \frac{1}{a}\sum_{i=1}^{a} \alpha_i,\ \bar{\beta} = \frac{1}{p}\sum_{j=1}^{p} \beta_j,\ \bar{\gamma} = \frac{1}{(a+p-1)}\sum_{k=1}^{a+p-1} \gamma_k.$$

Then, as with all fixed-effect ANOVA-type models, it is natural to reparameterize model (1) to its equivalent form

$$Y_{ij} = \mu^* + \alpha_i^* + \beta_j^* + \gamma_{a-i+j}^* + \varepsilon_{ij}, \tag{3}$$

where $\mu^* = (\mu + \bar{\alpha} + \bar{\beta} + \bar{\gamma})$, $\alpha_i^* = (\alpha_i - \bar{\alpha})$,

$$\beta_j^* = (\beta_j - \bar{\beta}), \text{ and } \gamma_{a-i+j}^* = (\gamma_{a-i+j} - \bar{\gamma});$$

clearly, then, we have

$$\sum_{i=1}^{a} \alpha_i^* = \sum_{j=1}^{p} \beta_j^* = \sum_{k=1}^{a+p-1} \gamma_k^* = 0. \tag{4}$$

It is important to emphasize that the reparameterized model (3) simply re-expresses each effect in model (1) as a deviation from the mean of all effects of that type, and such centering creates no distortion with respect to assessing patterns in estimated effects. In contrast, the use [22, 32] of unnatural constraints like $\alpha_1 = \beta_1 = \gamma_1 = 0$ does not lead to a straightforward equivalent representation of equation (1), and can produce misleading patterns in estimated coefficients.

The restrictions (4) imply, for example, that only the first $(a-1)$ age effects, the first $(p-1)$ period effects, and the first $(a+p-2)$ cohort effects in model (3) require estimation, since

$$\alpha_a^* = -\sum_{i=1}^{a-1} \alpha_i^*,\ \beta_p^* = -\sum_{j=1}^{p-1} \beta_j^*$$

and

$$\gamma_{a+p-1}^* = -\sum_{k=1}^{a+p-2} \gamma_k^*.$$

Employing (without loss in generality) these three equalities defines the $(ap) \times [2(a+p)-3]$ design matrix $\mathbf{X}^*$ for the matrix representation of model (3), namely

$$\mathcal{E}(\mathbf{Y}) = \mathbf{X}^*\boldsymbol{\xi}^*, \tag{5}$$

where the response vector is

and the paramet

Appendix A give
Table 2.

The identificati
full rank (i.e. th
interested reader
dependency for t

Assuming for
least squares, the
form

Since $\mathbf{X}^{*\prime}\mathbf{X}^*$ is o
necessary to imp
a unique solutio
be made carefull
lation linear fun

A common p
adjacent age, pe
$\hat{\alpha}_1^* = \hat{\alpha}_2^*$), such a
data-based impr
effects. However
a major impact
coefficients. As
estimated patte
function of the
any meaningful

An example
With $Y_{ij} = \ln(1$
based on etiolo
$\hat{\alpha}_2^*$ for the first
the linear const
constraint is th
these first two
in question; in
As we will dem
equations (6)
parameter vect

When the c
parameters are
the pattern in
Tables 3a and
age, a pattern
risk increases
constraints $\hat{\beta}_1^*$
fluctuations in

*Poisson regressi
its use is the
the true unde
fitting models
GLIM [35].

$$\mathbf{Y}' = (Y_{11}, \ldots, Y_{1p}; \; Y_{21}, \ldots, Y_{2p}; \ldots; Y_{a1}, \ldots, Y_{ap})$$

and the parameter vector is

$$\xi^{*\prime} = (\mu^*; \; \alpha_1^*, \ldots, \alpha_{a-1}^*; \; \beta_1^*, \ldots \; \beta_{p-1}^*; \; \gamma_1^*, \ldots, \gamma_{a+p-2}^*).$$

Appendix A gives the exact form of $\mathbf{X}^*$ for the special case $a = 3$, $p = 4$ diagrammed in Table 2.

The identification problem arises under model (5) because the matrix $\mathbf{X}^*$ is one less than full rank (i.e. there exists an exact linear dependency among the columns of $\mathbf{X}^*$). The interested reader can consult Appendix A to see the exact structure of this linear dependency for the special case $a = 3$, $p = 4$, and also for general a and p.

Assuming for now that the elements in ξ^* are to be estimated by ordinary (unweighted) least squares, the normal equations to be solved for the vector $\hat{\xi}^*$ of estimates are of the form

$$\mathbf{X}^{*\prime}\mathbf{X}^*\hat{\xi}^* = \mathbf{X}^{*\prime}\mathbf{Y}. \tag{6}$$

Since $\mathbf{X}^{*\prime}\mathbf{X}^*$ is one less than full rank, so that its inverse $(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}$ does not exist, it is necessary to impose at least one linear constraint on the elements of $\hat{\xi}^*$ in order to obtain a unique solution to the set of equations (6). (The choice of this linear constraint must be made carefully; in particular, statistical theory dictates that the corresponding population linear function of the elements of ξ^* must be non-estimable [33].)

A common practice in the literature has been to require that the estimates of two adjacent age, period, or cohort effects be numerically equal to one another (e.g. that $\hat{\alpha}_1^* = \hat{\alpha}_2^*$), such a requirement supposedly stemming either from *a priori* conjectures and/or data-based impressions about the true underlying age, period, and cohort population effects. However, the choice of such a constraint on the parameter estimates generally has a major impact on the observed patterns in the estimated age, period, and cohort coefficients. As has been illustrated both with real data [17] and with artificial data [20,22], estimated patterns in age, period, and cohort effects typically vary dramatically as a function of the constraint employed, such variation often being so extreme as to prohibit any meaningful interpretation of the data under consideration.

An example using the lung cancer data in Table 1 clearly illustrates this phenomenon. With $Y_{ij} = \ln(10^5 \hat{R}_{ij})$ for these data, a seemingly very reasonable choice for a constraint based on etiologic considerations would be to require that the estimated age effects $\hat{\alpha}_1^*$ and $\hat{\alpha}_2^*$ for the first two age groups (namely, 30–34 and 35–39) be equal; i.e. we are imposing the linear constraint $(\hat{\alpha}_1^* - \hat{\alpha}_2^*) = 0$. The underlying rationale for choosing this particular constraint is that it appears reasonable to assume that the young ages encompassed by these first two age groups would exhibit similar effects regarding the rare chronic disease in question; in other words, we are, in actuality, assuming that $\alpha_1^* = \alpha_2^*$ in the population. As we will demonstrate, if $\alpha_1^* \neq \alpha_2^*$, then the vector $\hat{\xi}^*$ obtained by solving the system of equations (6) under the constraint $\hat{\alpha}_1^* = \hat{\alpha}_2^*$ will *not* be an unbiased estimator of the parameter vector ξ^*.

When the constraint $(\hat{\alpha}_1^* - \hat{\alpha}_2^*) = 0$ is used in conjunction with equation (5), and the parameters are estimated either by least squares via equation (6) or by Poisson regression*, the pattern in the estimated age effects is in the opposite direction to that expected (see Tables 3a and b). In particular, the estimated age effects decrease in value with increasing age, a pattern which is diametrically opposed to the well-documented principle that cancer risk increases with age. For comparison purposes, the results based on using the constraints $\hat{\beta}_1^* = \hat{\beta}_2^*$ and $\hat{\gamma}_1^* = \hat{\gamma}_2^*$ are also presented in Tables 3a and b. The large fluctuations in the values of these estimates as a function of the chosen constraint are most

*Poisson regression is a maximum likelihood-based parameter estimation procedure. The theoretical basis for its use is the assumption (in our notation) that O_{ij} has a Poisson distribution with mean $N_{ij}\lambda_{ij}$, with λ_{ij} the true underlying rate in cell (i, j) and $\hat{R}_{ij}$ the estimator of λ_{ij}. Two popular computer programs for fitting models like $\ln(\lambda_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{a-i+j}$ by Poisson regression methods are CATMAX [34] and GLIM [35]. For an excellent discussion of Poisson regression methodology, see Frome [36].

disturbing, especially when faced with the fact that the constraints used in Tables 3a and b are not any more unrealistic than other possible constraint specifications that could be made.

With regard to the goodness of fit measures (i.e. R^2 and the deviance) given in Tables 3a and b, it is typically the case when fitting by least squares the multiple classification model (3) to APC data that R^2 values extremely close to 1 are obtained. However, this only means that the fitted model leads to excellent agreement between observed (Y_{ij}) and predicted ($\hat{Y}_{ij} = \hat{\mu}^* + \hat{\alpha}_i^* + \hat{\beta}_j^* + \hat{\gamma}_{a-i+j}^*$) responses; such agreement promises nothing about the accuracy of the *individual* estimated effects constituting the conglomerate estimate $\hat{Y}_{ij}$.

TABLE 3a. ESTIMATED MODEL (3) EFFECTS AS A FUNCTION OF CONSTRAINT CHOICE; METHOD OF ESTIMATION IS UNWEIGHTED LEAST SQUARES†; DATA ANALYZED ARE IN TABLE 1

$(\hat{\alpha}_1^* - \hat{\alpha}_2^* = 0$			$(\hat{\beta}_1^* - \hat{\beta}_2^* = 0$			$(\hat{\gamma}_1^* - \hat{\gamma}_2^* = 0$		
Age	Period	Cohort	Age	Period	Cohort	Age	Period	Cohort
1.424	−4.484	6.776	−4.690	0.406	−4.225	−1.334	−2.271	1.796
1.424	−3.262	6.223	−3.466	0.406	−3.557	−0.790	−1.601	1.796
1.372	−2.191	5.521	−2.296	0.254	−3.036	−0.288	−1.084	1.647
1.236	−1.038	4.846	−1.209	0.184	−2.489	0.129	−0.485	1.525
0.976	0.039	4.193	−0.246	0.039	−1.920	0.423	0.039	1.426
0.580	1.123	3.485	0.580	−0.099	−1.405	0.580	0.570	1.271
0.064	2.190	2.800	1.287	−0.255	−0.867	0.618	1.083	1.140
−0.575	3.280	2.108	1.870	−0.387	−0.337	0.532	1.620	1.001
−1.327	4.343	1.354	2.341	−0.548	0.132	0.334	2.129	0.801
−2.115		0.556	2.775		0.556	0.098		0.556
−3.059		−0.310	3.054		0.912	−0.292		0.243
		−1.236			1.209			−0.129
		−2.195			1.472			−0.535
		−3.194			1.696			−0.980
		−4.165			1.947			−1.398
		−5.048			2.286			−1.728
		−6.054			2.503			−2.180
		−7.157			2.623			−2.730
		−8.503			2.500			−3.522

†$R^2 = 0.999$ for each of these three sets of estimates.

TABLE 3b. ESTIMATED MODEL (3) EFFECTS AS A FUNCTON OF CONSTRAINT CHOICE; METHOD OF ESTIMATION IS POISSON REGRESSION (WLSIE)†; DATA ANALYZED ARE IN TABLE 1‡

$(\hat{\alpha}_1^* - \hat{\alpha}_2^* = 0$			$(\hat{\beta}_1^* - \hat{\beta}_2^* = 0$			$(\hat{\gamma}_1^* - \hat{\gamma}_2^* = 0$		
Age	Period	Cohort	Age	Period	Cohort	Age	Period	Cohort
1.549	−4.685	7.164	−4.890	0.466	−4.426	−1.415	−2.314	1.828
1.549	−3.397	6.571	−3.602	0.466	−3.732	−0.822	−1.619	1.828
1.492	−2.267	5.824	−2.372	0.309	−3.192	−0.287	−1.081	1.673
1.325	−1.064	5.066	−1.251	0.224	−2.662	0.139	−0.471	1.508
1.032	0.073	4.366	−0.256	0.073	−2.073	0.439	0.073	1.401
0.598	1.188	3.620	0.598	−0.099	−1.532	0.598	0.596	1.248
0.057	2.278	2.883	1.345	−0.298	−0.980	0.650	1.092	1.105
−0.627	3.394	2.153	1.949	−0.470	−0.422	0.559	1.615	0.968
−1.429	4.480	1.362	2.435	−0.671	0.074	0.350	2.109	0.769
−2.287		0.517	2.865		0.517	0.085		0.517
−3.259		−0.400	3.179		0.888	−0.296		0.193
		−1.372			1.204			−0.186
		−2.362			1.501			−0.583
		−3.349			1.802			−0.978
		−4.340			2.099			−1.376
		−5.246			2.481			−1.689
		−6.288			2.728			−2.137
		−7.389			2.914			−2.646
		−8.780			2.811			−3.443

†Poisson regression estimates are produced via a maximum likelihood-based procedure involving iteration from a set of initial estimates to a set of final estimates; the acronym WLSIE pertains to the fact that "weighted least squares initial estimates" were used to start the CATMAX [34] iteration procedure.

‡The standard goodness-of-fit statistic for Poisson regression is the *deviance*, which is approximately equal to the "observed versus predicted" statistic.

$$\sum_{i=1}^{a}\sum_{j=1}^{b}(O_{ij} - \hat{O}_{ij})^2/\hat{O}_{ij},$$

where $\hat{O}_{ij}$ is the predicted count in cell (i, j) using the fitted model. For each of the three sets of estimates above, the deviance value was 99.682; this suggests some lack of fit when compared to χ^2 tables with 63 *df*.

TABLE 3c. ESTIMATED MO

$(\hat{\alpha}_1^* - \hat{\alpha}_2^*)$	
Age	Perio
−2.855	−0.1
−2.855	−0.0
−1.018	−0.0
−0.277	−0.0
0.307	−0.0
0.731	0.0
1.009	0.0
1.136	1.1
1.099	0.1
1.010	
1.713	

†The acronym RSIE per procedure. In this exa

$i = 1,2,\ldots,a$; $\hat{\beta}_j^* = 0$. means and "null" effe

‡For each of the three

As mentioned observed (O_{ij}) a essentially equal criticized by son when large popu of an alternativ structure to R^2

One disturbi

is that it is not example, the O (calculated as 5-year *totals* o mately five tim value and each

Finally, a cc phenomenon. estimates obtai to generate th the values of th for the discrep for the iteratic surface for th maximum clos 3c). Although strong possibi

TABLE 3c. ESTIMATED MODEL (3) EFFECTS AS A FUNCTION OF CONSTRAINT CHOICE; METHOD OF ESTIMATION IS POISSON REGRESSION (RSIE)†; DATA ANALYZED ARE IN TABLE 1‡

$(\hat{\alpha}_1^* - \hat{\alpha}_2^* = 0$			$(\hat{\beta}_1^* - \hat{\beta}_2^* = 0$			$(\hat{\gamma}_1^* - \hat{\gamma}_2^* = 0$		
Age	Period	Cohort	Age	Period	Cohort	Age	Period	Cohort
−2.855	−0.131	0.270	−2.998	−0.012	−0.015	−2.973	−0.032	0.031
−2.855	−0.099	0.233	−1.996	−0.012	−0.010	−1.975	−0.027	0.031
−1.018	−0.069	0.201	−1.105	−0.011	−0.007	−1.090	−0.021	0.029
−0.277	−0.036	0.172	−0.335	−0.008	−0.003	−0.325	−0.014	0.027
0.307	−0.003	0.144	0.278	−0.005	0.000	0.283	−0.005	0.025
0.731	0.031	0.116	0.731	0.001	0.002	0.731	0.006	0.023
1.009	0.066	0.089	1.038	0.007	0.005	1.033	0.017	0.021
1.136	1.103	0.062	1.195	0.016	0.009	1.184	0.031	0.019
1.099	0.138	0.036	1.187	0.024	0.012	1.172	0.045	0.018
1.010		0.010	1.128		0.014	1.107		0.014
1.713		−0.019	0.877		0.014	0.853		0.009
		−0.051			0.011			0.000
		−0.084			0.007			−0.008
		−0.116			0.004			−0.017
		−0.148			0.000			−0.025
		−0.177			0.000			−0.031
		−0.210			−0.005			−0.041
		−0.238			−0.012			−0.053
		−0.290			−0.026			−0.072

†The acronym RSIE pertains to the fact that "researcher-selected initial estimates" were used to start the CATMAX [34] iteration procedure. In this example, the initial estimates were as follows:

$$\hat{\alpha}_i^* = \frac{1}{p}\sum_{j=1}^{p} \ln(10^5 \hat{R}_{ij}) - \frac{1}{ap}\sum_{i=1}^{a}\sum_{j=1}^{p} \ln(10^5 \hat{R}_{ij}),$$

$i = 1,2,\ldots,a$; $\hat{\beta}_j^* = 0$, $j = 1,2,\ldots,p$; $\hat{\gamma}_k^* = 0$, $k = 1,2,\ldots,a+p-1$. Comparable initial estimates (e.g. age-specific marginal means and "null" effects for period and cohort) have been used by other researchers [3, 12].

‡For each of the three sets of estimates above, the deviance value was larger than 10^6.

As mentioned in a footnote to Table 3b, the deviance reflects discrepancies between observed (O_{ij}) and predicted ($\hat{O}_{ij}$) cell counts, and hence can be large even when R^2 is essentially equal to 1 in value (as can be seen in Tables 3a and b). The deviance has been criticized by some researchers [3] as being too sensitive to departures from the fitted model when large populations at risk are under study. These researchers have suggested the use of an alternative measure of fit for such count data; this *ad hoc* statistic is similar in structure to R^2 and generally produces comparable values.

One disturbing, but not typically appreciated, characteristic of the deviance measure

$$\sum_{i=1}^{a}\sum_{j=1}^{p} (O_{ij} - \hat{O}_{ij})^2/\hat{O}_{ij}$$

is that it is not invariant with respect to scale changes in O_{ij} (and hence in $\hat{O}_{ij}$). As an example, the O_{ij} values in Table 1 are averages per year, so that the person-year values (calculated as $N_{ij} = O_{ij}/\hat{R}_{ij}$) are also on a per-year basis. If we had worked instead with 5-year *totals* of deaths and person-years, the deviance value would have been approximately five times larger than that given in Table 3b. In contrast, a scaling up of each O_{ij} value and each N_{ij} value by a factor of 5 leaves $\hat{R}_{ij}$, and hence R^2, unaffected.

Finally, a comparison of the results in Tables 3b and c highlights another alarming phenomenon. In particular, it is an unfortunate circumstance that the final parameter estimates obtained via an iterative maximum likelihood computer program (e.g. as used to generate the Poisson regression analysis results given in Tables 3b and c) depend on the values of the initial estimates used to start the iteration process. The underlying reason for the discrepancies among sets of parameter estimates based on different starting values for the iterations is that the iteration process often does not search the entire likelihood surface for the desired "global" maximum; instead, the iteration stops at a "local" maximum close to the set of initial estimates (which is apparently what happened in Table 3c). Although this undesirable phenomenon has been mentioned in the literature [3], the strong possibility of obtaining incorrect parameter estimates based on the use of such

likelihood surface search routines does not seem to have detracted from the popularity of such iteration algorithms.

Returning to our discussion of the results in Tables 3a and b, it can be shown that *estimation bias* is the primary reason why patterns in estimated age, period, and cohort effects vary so much as a function of the choice of the additional linear constraint which must be placed on the elements of $\hat{\xi}^*$ in order to fit model (5). A Searle [33] points out, the elements of $\hat{\xi}^*$ will be unbiased estimators of the corresponding elements of ξ^* only if the chosen constraint actually holds among the true (but unknown) population parameter values. When the chosen constraint does not hold *perfectly* in the population (a situation which almost always exists), the extent of the estimation bias attendant with the use of any linear constraint, like $(\hat{\alpha}_1^* - \hat{\alpha}_2^*) = 0$, will depend on how much the corresponding population linear function, like $(\alpha_1^* - \alpha_2^*)$, differs from zero.

Theorem 3.2 in Kupper *et al.* [28] characterizes the exact structure of the bias in $\hat{\xi}^*$ as a function of the "amount" by which the chosen constraint is in conflict with the true (unknown) population structure. Appendix B discusses this theorem and provides an example illustrating its use. The important general implication of Theorem 3.2 [28] is that the choice of constraint is *the* crucial determinant of the accuracy in the estimated age, period, and cohort effects based on model (5).

As mentioned earlier, APC researchers have employed two general strategies when searching for an appropriate constraint. The first method involves examining several sets of estimated coefficients, each set obtained via the use of a different constraint supposedly chosen based on *a priori* suppositions about the population under study. The typical impression gained from such an examination is that there is considerable variation in estimated coefficient values from set to set, and that a definitive choice as to which set of coefficients (if any) is most reliable is often difficult, if not impossible, to make.

Another popular approach for choosing a constraint involves a preliminary descriptive examination of patterns in the data to be analyzed. This data-based method utilizes certain observed trends (e.g. in the crude age-specific means) to suggest a possibly realistic constraint or to start a maximum likelihood iteration process for parameter estimation [3, 12]. Kupper *et al.* [28] discuss in theoretical terms why such data-dependent procedures can be quite misleading, and their position has been supported by Holford [7] and Rodgers [22, 23].

In theory, the following scenario describes about the only situation in which an APC analysis using model (5) might be informative. Suppose that several independent linear constraints on the elements of $\hat{\xi}^*$ can be specified, where each such constraint has strong justification on purely theoretical grounds (and not on hints obtained from a qualitative inspection of the observed data). If the separate sets of estimates obtained by applying each of these various theoretically-based constraints to model (5) are in quite close agreement with one another, then one might have some confidence in the reliability of this common set of estimated age, period, and cohort effects.

In reality, however, the above desirable situation very rarely occurs, and the APC data analyst is usually faced with choosing one particular set of estimated coefficients from several sets in which the coefficient estimates vary, often dramatically, from set to set. The final choice of one particular set of estimated coefficients (or, equivalently, of one particular constraint) is often a very subjective one, with preference given to estimated coefficient patterns which seem to support, at least approximately, certain *a priori* suspicions about relationships among the particular age, period, and cohort factors under study. Such subjective data-analysis decisions leave room for considerable controversy [22, 23, 25].

The estimated coefficients ultimately reported are often accompanied by their estimated "standard errors" [3, 37, 38]; such standard errors are available automatically as part of the computer output from popular model fitting programs (e.g. unweighted and weighted least squares, and Poisson regression packages). We hasten to discourage the use of such standard errors to construct confidence intervals for the age, period, and cohort effect parameters of interest under model (5). Such standard errors completely ignore the

potentially large
overwhelming s
of these varian
"closeness" of t
research would
coefficients usin
in estimated eff
study would aln
pursued in any

4.2. Two-factor

The discussio
of the paramet
cohort. One pc
model (5) is to
case a two-fact
of the data.

The use of s
recent years [3,
to demonstrate
(as we will pre
based on the
potentially in

In general, t
(5) is as follov
by least squar
these models (
regression) ar
fitting model (
fit is not signi
most appropr
to the data,
problem, wo

Kupper *et*
two-factor m
effects (age,
situation, me
out to be id
two-factor m
Without givi
phenomeno
regardless of
with respect
orientation

The follo
situation de
b, and the
presented in
obtained b

*It is importa
with respe
of *changes*
of the spe
values of
be used.

potentially large bias in the estimated coefficients. Since bias, and *not* variance, is the overwhelming source of discrepancy between estimated and true coefficient values, the use of these variance-based standard errors will create a misleading impression about the "closeness" of these estimates to the true population values. An area of needed statistical research would involve the development of realistic error bands for the estimation of APC coefficients using model (5); such error bands would have to take into account the bias in estimated effects due to inappropriate constraint specification. A detailed simulation study would almost certainly be needed, but such a difficult research effort has not yet been pursued in any productive way.

4.2. Two-factor models

The discussion thus far has focussed on statistical issues concerning accurate estimation of the parameters in model (5), a model involving all three of the factors age, period, and cohort. One possible way to avoid the identifiability problem attendant with the use of model (5) is to argue that one of the three factors (e.g. period) is "unimportant", in which case a two-factor (e.g. an age-cohort) model could reasonably provide a valid description of the data.

The use of such two-factor models to describe APC data has become quite popular in recent years [3, 10, 11, 17, 38, 39, 41, 42]. However, the statistical methodologies employed to demonstrate the "unimportance" of one of the three factors can be seriously misleading (as we will presently illustrate by example), and it is our position that published analyses based on the use of two-factor models (most typically of the age–cohort variety) are potentially in error.

In general, the strategy typically employed for choosing a two-factor model over model (5) is as follows. First of all, each of the three two-factor models is fit to the data (e.g. by least squares or Poisson regression methods). Next, measures of "goodness of fit" of these models (e.g. R^2 when using least squares, or the deviance [36] when using Poisson regression) are computed and then compared to the corresponding measure obtained by fitting model (5) to the data.* That two-factor model which best fits the data, and whose fit is not significantly different from that of the three-factor model, is then reported as the most appropriate model. (Of course, if none of the two-factor models provides a good fit to the data, then the three-factor model analysis, with its accompanying identifiability problem, would have to be pursued.)

Kupper *et al.* [28] have demonstrated both theoretically and by example that the above two-factor model selection strategy can be seriously misleading if one of the three sets of effects (age, period, or cohort) conforms to a true underlying *linear* pattern. In that situation, measures of fit like R^2 and the deviance for the three-factor model (5) will turn out to be identically equal in value to the R^2 and/or deviance values for that particular two-factor model *not* involving that specific set of effects satisfying a linear relationship. Without giving a formal mathematical argument, the underlying reason for this disturbing phenomenon is that the R^2 and deviance values based on fitting model (5) are the same regardless of the orientation of the linear pattern (since such measures of fit are invariant with respect to the choice of constraint, the determinant of that orientation), and one such orientation is the horizontal one indicating no non-zero effects for that factor.

The following numerical example provides a vivid illustration of the troublesome situation described above. The hypothetical data to be considered appear in Table 4a and b, and the results of various revealing Poisson regression analyses of these data are presented in Table 5. (Poisson regression-based parameter estimates were identical to those obtained by unweighted least squares, and so the latter results will not be reported.)

*It is important to note that measures of fit like R^2 and the deviance based on fitting model (5) are invariant with respect to the choice of constraint [33], so that the values of test statistics for assessing the significance of *changes* in R^2 and/or deviances in going from two-factor to three-factor models are the same regardless of the specific form of constraint used to fit (5). However, statistics [37, 40] which depend on the estimated values of individual coefficients under model (5), and hence vary with the choice of constraint, should not be used.

TABLE 4a. TABLE OF HYPOTHETICAL RATES PER 100,000 (i.e. $10^5 \hat{R}_{ij}$) UPON WHICH THE RESULTS IN TABLE 5 ARE BASED

		Period group (j)				
		$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
Age group (i)	$i=1$	1.73	1.82	1.82	1.73	1.65
	$i=2$	1.97	2.12	2.23	2.23	2.21
	$i=3$	2.10	2.39	2.56	2.69	2.69
	$i=4$	2.17	2.52	2.87	3.08	3.24
	$i=5$	2.08	2.53	2.94	3.35	3.60
	$i=6$	1.79	2.29	2.80	3.25	3.71
	$i=7$	1.51	1.94	2.50	3.05	3.54

TABLE 4b. TABLE OF HYPOTHETICAL VALUES OF O_{ij} UPON WHICH THE RESULTS IN TABLE 5 ARE BASED†

		Period group (j)				
		$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
Age group (i)	$i=1$	8650	9100	9100	8650	8250
	$i=2$	9850	10,600	11,150	11,150	10,600
	$i=3$	10,500	11,950	25,600	26,900	26,900
	$i=4$	10,850	25,200	28,700	30,800	32,400
	$i=5$	10,400	25,300	29,400	33,500	36,000
	$i=6$	8950	11,450	28,000	32,500	37,100
	$i=7$	7550	9700	12,500	30,500	35,400

†The hypothetical values of N_{ij} used in the Poisson regression analyses of Table 5 are obtained from the appropriate entries in Tables 4a and b via the formula $N_{ij} = O_{ij}/\hat{R}_{ij}$.

The rates in Table 4a were generated via model (3) as

$$\ln(10^5 \hat{R}_{ij}) = \mu^* + \alpha_i^* + \beta_j^* + \gamma_{7-i+j}^*,$$

where $\mu^* = (\bar{\alpha} + \bar{\beta} + \bar{\gamma})$; the values of these parameters are listed in the first column of Table 5 (for simplicity, we have assumed that $\mu = 0$ and that $\varepsilon_{ij} = 0$). As an example, when $i = j = 1$, then

$$\ln(10^5 \hat{R}_{ij}) = (0.220 + 0.300 + 0.280) \; -0.220 \; -0.200 \; +0.170 = 0.550$$

so that

$$10^5 \hat{R}_{ij} = e^{0.550} = 1.73$$

The entries in Table 5 illustrate numerically the following principles: (i) when fitting model (3), the use of a constraint (namely, $\hat{\alpha}_5^* = \hat{\alpha}_6^*$) which actually holds in the population (namely, $\alpha_5^* = \alpha_6^*$) yields unbiased estimators of the age, period, and cohort population effects (cf. columns {1} and {2}); (ii) the bias in the parameter estimators using model (3) varies considerably with the choice of constraint (cf. columns {2}, {3} and {4}); (iii) the deviance and R^2 statistics based on fitting model (3) are invariant with respect to the choice of constraint employed (cf. deviance and R^2 values for columns {2}, {3}, and {4}); and, (iv) when one set of population effects follows a straight line pattern (in this example, it is the set of period effects), then equating any two such estimated effects under model (3) "rotates" the (estimated) linear pattern to a horizontal position, creating the false impression that such a factor is "unimportant" (see column {3}).

With regard to principle (iv), an examination of the three two-factor model sets of estimates and their deviance and R^2 values (columns {5}, {6}, and {7}) leads to the conclusion that the fitted age–cohort (A–C) model would be chosen as the "best" model (along with its severely biased set of effect estimates). What this means in practice is that a non-significant improvement in fit over a two-factor model when using model (3) does *not* allow one to distinguish between the situation where there are really no non-zero population effects for the added third factor (i.e. the horizontal linear orientation) and the situation where the third factor population effects follow a *non*-horizontal linear relationship (thus suggesting an important role for that factor). If the latter situation actually exists, then the chosen two-factor model will produce a very misleading impression of the true patterns in the corresponding age, period, and cohort population effects (again, see

TABLE 5. AN ILLUSTRAT
PERIOD, AND COHORT EF

True APC parameter values (Col. {1})
Age effects ($\bar{\alpha} = 0.220$)
$\alpha_1^* = -0.220$
$\alpha_2^* = -0.120$
$\alpha_3^* = -0.030$
$\alpha_4^* = 0.055$
$\alpha_5^* = 0.110$
$\alpha_6^* = 0.110$
$\alpha_7^* = 0.095$
Period effects ($\bar{\beta} = 0.3$
$\beta_1^* = -0.200$
$\beta_2^* = -0.100$
$\beta_3^* = 0$
$\beta_4^* = 0.100$
$\beta_5^* = 0.200$
Cohort effects ($\bar{\gamma} = 0.2$
$\gamma_1^* = -0.280$
$\gamma_2^* = -0.130$
$\gamma_3^* = 0.020$
$\gamma_4^* = 0.120$
$\gamma_5^* = 0.170$
$\gamma_6^* = 0.200$
$\gamma_7^* = 0.170$
$\gamma_8^* = 0.120$
$\gamma_9^* = 0.020$
$\gamma_{10}^* = -0.130$
$\gamma_{11}^* = -0.280$
Deviance:
df:
R^2 values:

columns {1}, {
procedures wh
produce misle

The present
popular multi
use of model
interpretation
with hypothet
The strateg
two-factor mc
(age, period,
linear. In part
(i.e. there is g
or the devian
have been es
disturbing fa
seriously bias
Based on t
are exceeding
fit is of the m
for producin
way to avoid
in basic struc

TABLE 5. AN ILLUSTRATION OF THE FLUCTUATION IN NUMERICAL ESTIMATES, OBTAINED VIA POISSON REGRESSION [34], OF THE AGE, PERIOD, AND COHORT EFFECTS IN MODEL (3) AND IN CERTAIN TWO-FACTOR MODELS AS A FUNCTION OF THE CHOICE OF CONSTRAINT; $\mu = 0$ AND $\varepsilon_{ij} = 0$ FOR SIMPLICITY; DATA ANALYZED ARE IN TABLES 4a AND b

True APC parameter values (Col. {1})	Parameter estimates: Full APC model (3) $\hat{\alpha}_5^* = \hat{\alpha}_6^*$ Col. {2}	$\hat{\beta}_1^* = \hat{\beta}_2^*$ Col. {3}	$\hat{\gamma}_8^* = \hat{\gamma}_7^*$ Col. {4}	Two-factor models: A–P Col. {5}	A–C Col. {6}	P–C Col. {7}
Age effects ($\bar{\alpha} = 0.220$)						
$\alpha_1^* = -0.220$	−0.220	−0.520	−0.310	−0.340	−0.520	
$\alpha_2^* = -0.120$	−0.120	−0.320	−0.180	−0.142	−0.320	
$\alpha_3^* = -0.030$	−0.030	−0.130	−0.060	−0.007	−0.130	
$\alpha_4^* = 0.055$	0.055	0.055	0.055	0.119	0.055	
$\alpha_5^* = 0.110$	0.110	0.210	0.140	0.170	0.210	
$\alpha_6^* = 0.110$	0.110	0.310	0.170	0.140	0.310	
$\alpha_7^* = 0.095$	0.095	0.395	0.185	0.060	0.395	
Period effects ($\bar{\beta} = 0.300$)						
$\beta_1^* = -0.200$	−0.200	0	−0.140	−0.255		−0.297
$\beta_2^* = -0.100$	−0.100	0	−0.070	−0.100		−0.144
$\beta_3^* = 0$	0	0	0	0.030		−0.001
$\beta_4^* = 0.100$	0.100	0	0.070	0.128		0.144
$\beta_5^* = 0.200$	0.200	0	0.140	0.197		0.298
Cohort effects ($\bar{\gamma} = 0.280$)						
$\gamma_1^* = -0.280$	−0.280	−0.780	−0.430		−0.780	−0.082
$\gamma_2^* = -0.130$	−0.130	−0.530	−0.250		−0.530	0.050
$\gamma_3^* = 0.020$	0.020	−0.280	−0.070		−0.280	0.178
$\gamma_4^* = 0.120$	0.120	−0.080	0.060		−0.080	0.235
$\gamma_5^* = 0.170$	0.170	0.070	0.140		0.070	0.241
$\gamma_6^* = 0.200$	0.200	0.200	0.200		0.200	0.243
$\gamma_7^* = 0.170$	0.170	0.270	0.200		0.270	0.162
$\gamma_8^* = 0.120$	0.120	0.320	0.180		0.320	0.049
$\gamma_9^* = 0.020$	0.020	0.320	0.110		0.320	−0.126
$\gamma_{10}^* = -0.130$	−0.130	0.270	−0.010		0.270	−0.361
$\gamma_{11}^* = -0.280$	−0.280	0.220	−0.130		0.220	−0.589
Deviance:	0.483	0.483	0.483	5329	0.483	1041
df:	15	15	15	24	18	20
R^2 values:	1	1	1	0.985	1	0.998

columns {1}, {3}, and {6} in Table 5). These disturbing results suggest that APC analysis procedures which hinge on the use of two-factor models [10, 11] have the potential to produce misleading conclusions.

5. DISCUSSION

The presentation in the previous sections focussed on statistical APC analysis using the popular multiple classification model (3). The identifiability problem attendant with the use of model (3) was discussed theoretically, and its deleterious impact upon the interpretation of APC data analysis efforts was illustrated numerically both with real and with hypothetical data.

The strategy of circumventing this identifiability problem via the use of a "good-fitting" two-factor model was shown to be potentially misleading when one of the three factors (age, period, or cohort) has an underlying population effect pattern which is essentially linear. In particular, the fact that a two-factor or three-factor APC model fits the data well (i.e. there is good agreement between observed and predicted *responses* as measured by R^2 or the deviance) does *not* allow one to conclude that individual coefficients in that model have been estimated with validity. Indeed, an inspection of Table 5 reinforces the disturbing fact that good fitting APC models can involve estimated effects which are seriously biased.

Based on the above considerations, it is our opinion that valid statistical APC analyses are exceedingly difficult, if not impossible, to carry out when the underlying model to be fit is of the multiple classification form (3). What, then, are some viable approaches, if any, for producing a valid statistical modeling analysis of APC data? Certainly, the most direct way to avoid the identifiability problem is to consider using a model form which is different in basic structure from that of model (3). Various approaches involving alternative model

forms have been considered in the literature [2, 4, 6, 7, 11, 12, 41], some of which may be "different enough" from model (3) to avoid the identifiability problem. However, none of these approaches or models has received anywhere near the statistical scrutiny given to the multiple classification model (3). Until comparable evaluations are made about these alternative methodologies, judgment must be withheld regarding their utility for producing valid analyses of APC data.

The modeling controversy aside, an even more basic problem with APC analyses concerns the characteristics of the data used in such analyses. Firstly, it is well known that the accuracy of time-related patterns in mortality and morbidity rates can be severely compromised by a number of possible sources of error (e.g. diagnostic errors, changes in diagnostic procedures over time, errors in determining person-years at risk, etc.). In modeling trends in such rates over time as a function of the age, period, and cohort factors, it is necessary to realize that such modeling exercises (indeed, any statistical analyses) are futile if the data base itself is unreliable.

As mentioned quite early in this paper, one unique characteristic of APC data is that the number of rates associated with each of the various cohorts varies from 1 to min(a, p) as one moves from the extreme diagonals to the middle diagonal(s) of an age-by-period two-way layout (again, refer to Table 1). A re-casting of the data in Table 1 into an age-by-cohort two-way display (see Table 6) dramatically illustrates this variation in cohort-specific data. Loosely speaking, each empty cell in Table 6 can be looked upon as a "missing observation" in the sense that no cohort is observed over the entire range of age categories (nor over the entire range of period categories). Indeed, two birth cohorts in Table 6 (namely, the 1851 and 1941 cohorts) are associated with only one age group each, two cohorts (namely 1856 and 1936) involve only two age groups each, etc. Overall, 110 of the 209 cells in Table 6 (i.e. 52% of the cells) are empty.

Researchers have recognized this problem and have attempted to adjust for it in various ways (e.g. by weighting each cohort coefficient inversely with the number of observations specific to that cohort [27], or by discarding estimates of cohort effects involving small numbers of observations [3]). Such adjustments are basically *ad hoc* in nature, and do not justify extrapolation of estimated trends to future cohorts and future periods. As with any statistical modeling analysis, only prediction *within* the range of age, period, and cohort categories actually observed is permissible; and, even then, such interpolation is questionable when it involves those cohorts containing few observations.

In summary, the statistical analysis of APC data is plagued by many unresolved issues and potential sources of error. First of all, the selection of the response function $Y_{ij} = f(\hat{R}_{ij})$ to be modeled can have a significant influence on patterns in estimated age, period, and (especially) cohort effects. Moreover, even if specification of a meaningful function $f(\hat{R}_{ij})$

TABLE 6. LUNG CANCER MORTALITY DATA OF TABLE 1 ARRANGED BY CENTRAL YEAR OF BIRTH AND BY AGE GROUP AT DEATH

Central year of birth	Age group at death										
	30–34	35–39	40–44	45–49	50–54	55–59	60–64	65–69	70–74	75–79	80–84
1851											12.47
1856										19.09	23.49
1861									20.77	33.59	31.46
1866								20.73	32.98	44.96	64.03
1871							19.61	38.22	50.36	77.73	96.88
1876						16.14	34.75	54.29	85.62	115.21	132.16
1881					13.04	30.07	53.65	87.47	128.85	160.81	179.54
1886				8.19	21.76	47.06	87.38	136.24	184.99	224.70	260.91
1891			4.51	13.72	30.35	68.18	120.90	189.11	245.18	318.68	360.16
1896		2.38	6.78	17.62	43.32	88.37	158.32	240.25	322.38	415.37	
1901	1.08	3.06	8.46	22.31	53.82	108.86	189.36	289.07	390.94		
1906	1.43	3.66	10.05	27.46	64.29	121.86	218.81	325.13			
1911	1.64	4.10	11.75	30.75	72.11	139.46	238.48				
1916	1.58	4.32	13.40	36.16	81.43	151.86					
1921	1.52	5.26	16.07	41.06	86.42						
1926	1.96	6.35	19.60	47.51							
1931	2.16	7.25	19.87								
1936	2.14	7.21									
1941	1.74										

can be made, tl
the multiple cla
effect patterns.
assumptions wl
analysis. The ex
which almost c
accurately meas
for assessing th

The adoption
when the popu
non-horizontal
documenting (v
values) the exist
strategy in the s
if a fitted APC
for prediction (
the lack of coh

Given these p
modeling proce
provide import
possible that s
framework" [25
a priori assump
under study.

With regard
develop and ev
the identifiabili
period, and coh
taken place, we
analysis is still

Acknowledgements
regarding several of
the National Institu
Research—U.S.A..

1. Frost WH: T
1939
2. Moolgavkar S
Inst 62: 493-
3. Stevens RG,
4. Walter SD, M
incidence and
5. Barrett JC: A
1973
6. James IR, Seg
to prostate ca
7. Holford TR:
8. Moolgavkar
J Natl Cancer
9. Stevens RG,
and bladder.
10. Gardner MJ.
technique and
11. Osmond C, (
245–259, 198
12. Stevens RG,
119: 624–641
13. Van Der Hor

can be made, the choice of constraint on the elements of ξ^* in model (5) required to fit the multiple classification model (3) has an even more dramatic impact on such estimated effect patterns. Unfortunately, such a constraint choice must be based on *a priori* assumptions whose validity cannot be demonstrated empirically with the data under analysis. The extent of the bias in estimated effects resulting from the use of a constraint which almost certainly does not hold exactly in the population under study cannot be accurately measured; also, such bias invalidates the use of variance-based standard errors for assessing the "accuracy" of such estimated effects.

The adoption of a two-factor model based on standard goodness-of-fit criteria is invalid when the population effects for one of the factors (age, period, or cohort) follow a non-horizontal *linear* pattern. Hence, to attempt to avoid the identifiability problem by documenting (via statistical testing procedures based on changes in deviance and R^2 values) the existence of a "good-fitting" two-factor model can be a seriously misleading strategy in the sense that an important factor may be erroneously discounted. Finally, even if a fitted APC model can be found which "appears" reasonable, the use of such a model for prediction (e.g. about present or future cohort effect patterns) is severely limited by the lack of cohort-specific data.

Given these potential sources of error attendant with currently available APC statistical modeling procedures, it is our position that such regression methods cannot be said to provide important interpretational advantages over traditional graphical approaches. It is possible that such modeling procedures may constitute part of a general "accounting framework" [25, 43], but the potential for researcher bias is great because of the need for *a priori* assumptions concerning underlying population relationships among the variables under study.

With regard to future research efforts, we recommend that further work be done to develop and evaluate (in a rigorous statistical sense) innovative procedures which by-pass the identifiability problem, but which still provide *relevant* information regarding age, period, and cohort effect patterns. Since such development and evaluation has not yet taken place, we are forced to conclude that the current state-of-the-art of statistical APC analysis is still in its infancy.

Acknowledgements—We wish to thank Dr Ibrahim A. Salama and Dr Carl N. Yoshizawa for their insights regarding several of the issues addressed in this paper. We also wish to acknowledge training grant support from the National Institute of Environmental Health Sciences, and research support from The Council for Tobacco Research—U.S.A., Inc.

REFERENCES

1. Frost WH: The age selection of mortality from tuberculosis in successive decades. **Am J Hyg** 30: 91–96, 1939
2. Moolgavkar SH, Stevens RG, Lee JAH: Effect of age on incidence of breast cancer in females. **J Natl Cancer Inst** 62: 493–501, 1979
3. Stevens RG, Moolgavkar SH, Lee JAH: Temporal trends in breast cancer. **Am J Epid** 115: 759–777, 1982
4. Walter SD, Miller CT, Lee JAH: The use of age-specific mean cohort slopes in the analysis of epidemiologic incidence and mortality data. **J R Stat Soc A** 139: 227–245, 1976
5. Barrett JC: Age, time and cohort factors in mortality from cancer of the cervix. **J Hyg Camb** 71: 253–259, 1973
6. James IR, Segal MR: On a method of mortality analysis incorporating age-year interaction, with application to prostate cancer mortality. **Biometrics** 38: 433–443, 1982
7. Holford TR: The estimation of age, period and cohort effects for vital rates. **Biometrics** 39: 311–324, 1983
8. Moolgavkar SH and Stevens RG: Smoking and cancers of bladder and pancreas: risks and temporal trends. **J Natl Cancer Inst** 67: 15–23, 1981
9. Stevens RG, Moolgavkar SH: Estimation of relative risk from vital data: smoking and cancers of the lung and bladder. **J Natl Cancer Inst** 62: 1351–1357, 1979
10. Gardner MJ, Osmond C: Interpretation of disease time trends: is cancer on the increase? A simple cohort technique and its relationship to more advanced models. **J Epid Commun Health** 37: 274–278, 1983
11. Osmond C, Gardner MJ: Age, period, and cohort models applied to cancer mortality rates. **Stat Med** 1: 245–259, 1982
12. Stevens RG, Moolgavkar SH: A cohort analysis of lung cancer and smoking in British males. **Am J Epid** 119: 624–641, 1984
13. Van Der Hoff NM: Cohort analysis of lung cancer in the Netherlands. **Int J Epid** 8: 41–47, 1979

14. Kleinbaum DG, Kupper LL, Morgenstern H: **Epidemiologic Research: Principles and Quantitative Methods.** Belmont, California: Lifetime Learning Publications, 1982
15. Glenn ND: **Cohort Analysis.** No. 5 in the series: **Quantitative Applications in the Social Sciences.** Beverly Hills, CA: Sage Publications, 1977
16. Baltes PB: Longitudinal and cross-sectional sequences in the study of age and generation effects. **Hum Dev** 11: 145–171, 1968
17. Fienberg SE, Mason WM: Identification and estimation of age-period-cohort models in the analysis of discrete archival data. In **Sociological Methodology.** Schuessler KF (Ed.) San Francisco: Jossey-Bass, 1978
18. Honig M, Hanoch G: A general model of labor-market behavior of older persons. **Soc Secur Bull** 43: 29–39, 1980
19. Knoke D, Hout M: Social and demographic factors in American political party affiliation, 1952–72. **Am Soc Rev** 39: 700–713, 1974
20. Mason KO, Mason WM, Winsborough HH, Poole WK: Some methodological issues in cohort analysis of archival data. **Am Sociol Rev** 38: 242–258, 1973
21. Namboodiri NK: On factors affecting fertility at different stages in the reproduction history: an exercise in cohort analysis. **Social Forces** 59: 1114–1129, 1981
22. Rodgers WL: Estimable functions of age, period, and cohort effects. **Am Sociol Rev** 47: 774–787, 1982a
23. Rodgers WL: Reply to comment by Smith, Mason, and Fienberg. **Am Sociol Rev** 47: 793–796, 1982b
24. Schaie KW: A general model for the study of development problems. **Psychol Bull** 64: 92–107, 1965
25. Smith HL, Mason WM, Fienberg SE: More chimeras of the age–period–cohort accounting framework: comment on Rodgers. **Am Sociol Rev** 47: 787–793, 1982
26. Winsborough HH: Age, period, cohort, and education effects on earnings by race—an experiment with a sequence of cross-sectional surveys. In **Social Indicator Models.** Land KC, Spilleman S (Eds), New York: Russel Sage, 1975
27. Greenberg BG, Wright JJ, Sheps CG: A technique for analyzing some factors effecting the incidence of syphilis. **J Am Stat Assoc** 45: 373–399, 1950
28. Kupper LL, Janis JM, Salama IA, Yoshizawa CN, Greenberg BG: Age–period–cohort analysis: an illustration of the problems in assessing interaction in one observation per cell data. **Commun Stat** 12: 2779–2807, 1983
29. Springett VH: The beginning of the end of the increase in mortality from carcinoma of the lung. **Thorax** 21: 132–138, 1966
30. Susser M: Period effects, generation effects and age effects in peptic ulcer mortality. **J Chron Dis** 35: 29–40, 1982
31. Freeman OH, Holford TR: Summary rates. **Biometrics** 36: 195–205, 1980
32. Collins JJ: The contribution of medical measures to the decline of mortality from respiratory tuberculosis: an age–period–cohort model. **Demography** 19: 409–427, 1982
33. Searle SR: **Linear Models.** New York: Wiley, 1971
34. Stokes ME, Koch GG: A new macro for maximum likelihood fitting of long-linear models to Poisson and multinomial counts with contrast matrix capability for hypothesis testing. **Proceedings of the Eighth Annual SAS Users Group International Conference:** 795–800, 1983
35. Baker RJ, Nelder JA: **Generalized Linear Interactive Modeling (GLIM), Release 3.** Oxford: Numerical Algorithms Group, 1978
36. Frome EL: The analysis of rates using Poisson regression models. **Biometrics** 39: 665–674, 1983
37. Selvin S, Sacks S: Analysis of cohort effects from cross-sectional data. **Comp Prog Biomed** 9: 218–222, 1979
38. Stevens RG, Merkie EJ, Lustbader ED: Age and cohort effects in primary liver cancer. **Int J Cancer** 33: 453–458, 1984
39. Breslow NE, Day NE: Indirect standardization and multiplicative models for rates, with reference to the age adjustment of cancer incidence and relative frequency data. **J Chron Dis** 28: 289–303, 1975
40. Sacks ST, Selvin S: A method for detecting a cohort exposure. **Environ Res** 25: 167–177, 1981
41. Day NE, Charnay B: Time trends, cohort effects, and aging as influence on cancer incidence. In: Magnus K, ed., **Trends in Cancer Incidence (Causes and Practical Implications).** Magnus K (Ed.) Washington: Hemisphere Publishing Corporation, 1982
42. Gardner MJ, Osmond C: Interpretation of time trends in disease rates in the presence of generation effects. **Stat Med** 3: 113–130, 1984
43. Mason WM, Smith HL: Age–period–cohort analysis and the study of deaths from pulmonary tuberculosis. **Research Report 81–19,** Ann Arbor: The Population Studies Center of the University of Michigan. **Cohort Analysis.** Winsborough H, Duncan OD (Eds). New York: Academic Press, 1984

APPENDIX A

Consider model (5) for the special case $a = 3$, $p = 4$ diagrammed in Table 2. Then,

$$\mathbf{Y}' = (Y_{11}, Y_{12}, Y_{13}, Y_{14}; Y_{21}, Y_{22}, Y_{23}, Y_{24}; Y_{31}, Y_{32}, Y_{33}, Y_{34})$$

and

$$\xi^{*\prime} = (\mu^*; \alpha_1^*, \alpha_2^*; \beta_1^*, \beta_2^*, \beta_3^*; \gamma_1^*, \gamma_2^*, \gamma_3^*, \gamma_4^*, \gamma_5^*).$$

It thus follows that

$\mathbf{X}^*$

Although it is cer
rank; in other words
represent $\mathbf{X}^*$ by its

so that, for exampl
etc. Then, the interc

To check that equa
For general a an

Kupper *et al.* ([28],
the columns of the

It is easy to show
The form of equ
effect of an equally

for $i = 1, 2, \ldots, l,$
linear component o
reflects the fact tha
higher-order effect

Equation (A.3)

where $\mathbf{X}^*$ is given

$$\mathbf{v}' = \left[0; 1 - \frac{(a$$

Now, let

denote a set of e

based on using t
constants definin

then $\mathbf{c}'\xi_c^* = (\alpha_1^* -$
age effects to be
obtain the sets o

It thus follows that the (12×11) matrix $\mathbf{X}^*$ has the structure

$$\mathbf{X}^* = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 1 \\ 1 & -1 & -1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Although it is certainly not obvious based on a brief inspection, this matrix $\mathbf{X}^*$ is actually one less than full rank; in other words, there is an *exact* linear dependency among the columns of $\mathbf{X}^*$. To see this, let us equivalently represent $\mathbf{X}^*$ by its columns as

$$\mathbf{X}^* = (\mathbf{1};\ \mathbf{A}_1^*,\ \mathbf{A}_2^*;\ \mathbf{B}_1^*,\ \mathbf{B}_2^*,\ \mathbf{B}_3^*;\ \mathbf{C}_1^*,\ \mathbf{C}_2^*,\ \mathbf{C}_3^*,\ \mathbf{C}_4^*,\ \mathbf{C}_5^*),$$

so that, for example, $\mathbf{1}$ is the (12×1) column vector of ones, $\mathbf{A}_1^{*\prime} = (1, 1, 1, 1, 0, 0, 0, 0, -1, -1, -1, -1)$, etc. Then, the interested reader can verify that the following linear relationship holds among the columns of $\mathbf{X}^*$:

$$-\mathbf{A}_1^* + \tfrac{3}{2}\mathbf{B}_1^* + \tfrac{1}{2}\mathbf{B}_2^* - \tfrac{1}{2}\mathbf{B}_3^* - \tfrac{5}{2}\mathbf{C}_1^* - \tfrac{3}{2}\mathbf{C}_2^* - \tfrac{1}{2}\mathbf{C}_3^* + \tfrac{1}{2}\mathbf{C}_4^* + \tfrac{3}{2}\mathbf{C}_5^* = \mathbf{0}. \tag{A.1}$$

To check that equation (A.1) holds requires a row-by-row inspection of $\mathbf{X}^*$.

For general a and p, the column representation of $\mathbf{X}^*$ under model (5) is

$$\mathbf{X}^* = (\mathbf{1};\ \mathbf{A}_1^*, \ldots, \mathbf{A}_{a-1}^*;\ \mathbf{B}_1^*, \ldots, \mathbf{B}_{p-1}^*;\ \mathbf{C}_1^*, \ldots, \mathbf{C}_{a+p-2}^*). \tag{A.2}$$

Kupper *et al.* ([28], Theorem 3.1) have shown that the general expression for the linear constraint holding among the columns of the $(ap) \times [2(a+p)-3]$ matrix (A.2) is as follows:

$$\sum_{i=1}^{a-1}\left[i - \frac{(a+1)}{2}\right]\mathbf{A}_i^* - \sum_{j=1}^{p-1}\left[j - \frac{(p+1)}{2}\right]\mathbf{B}_j^* + \sum_{k=1}^{a+p-2}\left[k - \frac{(a+p)}{2}\right]\mathbf{C}_k^* = \mathbf{0}. \tag{A.3}$$

It is easy to show that equation (A.3) reduces to equation (A.1) when $a = 3$ and $p = 4$.

The form of equation (A.3) is revealing. Since the orthogonal polynomial coefficients for assessing the *linear* effect of an equally-spaced variable with l levels are given by the values of

$$\left\{i - \frac{(l+1)}{2}\right\}$$

for $i = 1, 2, \ldots, l$, equation (A.3) says, loosely speaking, that "(the linear component of the age columns) − (the linear component of the period columns) + (the linear component of the cohort columns) equals zero." This result reflects the fact that it is the *linear* effects of the age, period, and cohort categorical variables (in contrast to their higher-order effects) which are inextricably mixed, and hence are not individually identifiable.

APPENDIX B

Equation (A.3) in Appendix A can be compactly written in matrix notation as

$$\mathbf{X}^*\mathbf{v} = \mathbf{0}$$

where $\mathbf{X}^*$ is given by (A.2) and where

$$\mathbf{v}' = \left[0;\ 1 - \frac{(a+1)}{2},\ 2 - \frac{(a+1)}{2}, \ldots, (a-1) - \frac{(a+1)}{2};\ \frac{(p+1)}{2} - 1,\ \frac{(p+1)}{2} - 2, \ldots,\right.$$
$$\left.\frac{(p+1)}{2} - (p-1);\ 1 - \frac{(a+p)}{2},\ 2 - \frac{(a+p)}{2}, \ldots, (a+p-2) - \frac{(a+p)}{2}\right]. \tag{B.1}$$

Now, let

$$\hat{\xi}_c^{*\prime} = (\hat{\mu}^*;\ \hat{\alpha}_1^*, \ldots,\ \hat{\alpha}_{a-1}^*;\ \hat{\beta}_1^*, \ldots,\ \hat{\beta}_{p-1}^*;\ \hat{\gamma}_1^*, \ldots,\ \hat{\gamma}_{a+p-2}^*)$$

denote a set of estimates of the elements of

$$\xi^{*\prime} = (\mu^*;\ \alpha_1^*, \ldots, \alpha_{a-1}^*;\ \beta_1^*, \ldots, \beta_{p-1}^*;\ \gamma_1^*, \ldots, \gamma_{a+p-2}^*)$$

based on using the constraint $\mathbf{c}'\xi_c^* = 0$, where $\mathbf{c}'$ is an appropriately chosen $1 \times [2(a+p)-3]$ row vector of constants defining the constraint of interest. For example, when

$$\mathbf{c}' = (0,\ 1,\ -1,\ 0,\ 0, \ldots, 0),$$

then $\mathbf{c}'\hat{\xi}_c^* = (\hat{\alpha}_1^* - \hat{\alpha}_2^*) = 0$, so that $\hat{\xi}_c^*$ is that vector of estimates obtained by *forcing* the estimates of the first two age effects to be equal (i.e. by requiring that $\hat{\alpha}_1^* = \hat{\alpha}_2^*$). Recall that this particular constraint was employed to obtain the sets of estimates given in Tables 3a, b and c using the lung cancer mortality data in Table 1.

J Chron Dis Vol. 38, No.
Printed in Great Britain.

Kupper *et al.* ([28], Theorem 3.2) have shown that the bias in $\hat{\xi}_c^*$ when used to estimate ξ^* under model (5) has the specific structure

$$\text{Bias}(\hat{\xi}_c^*) = \mathcal{E}(\hat{\xi}_c^*) - \xi^* = k\mathbf{v} \tag{B.2}$$

where $\mathbf{v}$ is defined in equation (B.1) and where

$$k = -\mathbf{c}'\xi^*/\mathbf{c}'\mathbf{v}; \tag{B.3}$$

note that $\mathbf{c}'\mathbf{v} \neq 0$ since $\mathbf{c}'\xi^*$ is required to be non-estimable.

It is clear from equations (B.2) and (B.3) that the elements of $\hat{\xi}_c^*$ will all be unbiased estimators of the corresponding elements in ξ^* under model (5) when $k = 0$, or equivalently, when $\mathbf{c}'\xi^* = 0$. In other words, $\mathcal{E}(\hat{\xi}_c^*) = \xi^*$ when the constraint $\mathbf{c}'\hat{\xi}_c^* = 0$ employed to obtain the vector of estimates $\hat{\xi}_c^*$ actually holds in the population (i.e. $\mathbf{c}'\xi^* = 0$). [Under model (5),

$$\mathcal{E}(\hat{\alpha}_a^*) = -\sum_{i=1}^{a-1} \mathcal{E}(\hat{\alpha}_i^*),$$

with similar expressions holding for $\mathcal{E}(\hat{\beta}_p^*)$ and $\mathcal{E}(\hat{\gamma}_{a+p-1}^*)$. In general, $\mathbf{v}$, $\mathbf{c}$, and ξ^* change in definition depending on which particular age, period, and cohort parameters are removed using the set of restrictions (4), but the values of the biases based on equations (B.2) and (B.3) do not.]

As an example, consider again the constraint

$$\mathbf{c}'\hat{\xi}_c^* = (\hat{\alpha}_1^* - \hat{\alpha}_2^*),$$

where

$$\mathbf{c}' = (0,\ 1,\ -1,\ 0,\ 0, \ldots, 0).$$

Using $\mathbf{v}$ as defined in equation (B.1), we then have

$$\mathbf{c}'\mathbf{v} = \left[1 - \frac{(a+1)}{2}\right] - \left[2 - \frac{(a+1)}{2}\right] = -1$$

and

$$\mathbf{c}'\xi^* = (\alpha_1^* - \alpha_2^*),$$

so that $k = (\alpha_1^* - \alpha_2^*)$ and

$$\text{Bias}(\hat{\xi}_c^*) = (\alpha_1^* - \alpha_2^*)\mathbf{v}.$$

For example,

$$\text{Bias}(\hat{\beta}_1^*) = \mathcal{E}(\hat{\beta}_1^*) - \beta_1^* = (\alpha_1^* - \alpha_2^*)\left[\frac{(p+1)}{2} - 1\right].$$

Thus, unless an element of $\mathbf{v}$ is zero (in which case the parameter corresponding to that element will be unbiasedly estimated under the assumed model (5) regardless of the constraint $\mathbf{c}'\hat{\xi}^* = 0$ which is chosen), the value of the bias will be a function both of the sign and magnitude of the true (and unknown) parametric difference $(\alpha_1^* - \alpha_2^*)$. In particular, the bias in a particular effect estimator could be positive or negative depending on the (unknown) sign of $(\alpha_1^* - \alpha_2^*)$, and the magnitude of the bias will depend on the (unknown) value of $|\alpha_1^* - \alpha_2^*|$. It is no surprise, then, that different choices for a constraint can lead to widely different sets of estimated effect values (again, refer to the numerical results in Tables 3a, b and c).

Finally, expressions (B.2) and (B.3) lead to a very simple and informative condition for determining which linear functions of ξ^* are estimable (i.e. are estimated with no bias). In particular, if $\mathbf{l}'\xi^*$ is some linear function of ξ^* of interest [e.g. $\mathbf{l}'\xi^* = \alpha_1^*$ when $\mathbf{l}' = (0,\ 1,\ 0,\ 0, \ldots, 0)$], then it follows directly from equations (B.2) and (B.3) that

$$\text{Bias}(\mathbf{l}'\hat{\xi}_c^*) = \mathcal{E}(\mathbf{l}'\hat{\xi}_c^*) - \mathbf{l}'\xi^* = \mathbf{l}'\mathcal{E}(\hat{\xi}_c^*) - \mathbf{l}'\xi^* = \mathbf{l}'(k\mathbf{v}) = k\mathbf{l}'\mathbf{v},$$

so that

$$\text{Bias}(\mathbf{l}'\hat{\xi}_c^*) = 0 \text{ when } \mathbf{l}'\mathbf{v} = 0.$$

The condition $\mathbf{l}'\mathbf{v} = 0$ provides a very simple check for estimability, as opposed to the more complex partitioned matrix-based conditions given by Holford [7]. As a simple example, α_1^* is not estimable since, from equation (B.1),

$$\mathbf{l}'\mathbf{v} = (0,\ 1,\ 0,\ 0, \ldots, 0)\mathbf{v} = \left[1 - \frac{(a+1)}{2}\right] \neq 0, \quad a > 1.$$

More generally, since $\mathbf{v}$ in equation (B.1) involves the orthogonal polynomial coefficients for the *linear* effects of the age, period, and cohort factors, it follows that $\mathbf{l}'\mathbf{v} = 0$ *by definition* when $\mathbf{l}$ consists of the orthogonal polynomial coefficients for quadratic and higher-order effects.

The above simple proof that it is the non-linear effects of the age, period, and cohort factors which are estimable has been demonstrated via more complex arguments [17]. Holford [7] has made use of this result to estimate these non-linear effects using orthogonal polynomials. However, to relate his non-linear effect estimates to the parameters of interest in model (5) still necessitates making, as Holford acknowledges, a generally unverifiable assumption about the values of one or more parameters (i.e. the identifiability problem has *not* been avoided). As we have demonstrated, when such an assumption does not hold, there is strong potential for obtaining seriously biased estimates of the elements of ξ^* in model (5).